Apprentissage de SVM sur Données Bruitées

Guillaume Stempfel, Liva Ralaivola, François Denis

Laboratoire d'Informatique Fondamentale de Marseille, UMR CNRS 6166
Université de Provence, 39, rue Joliot Curie, 13013 Marseille, France
{guillaume.stempfel,liva.ralaivola,francois.denis}@lif.univ-mrs.fr

Abstract : Après avoir exhibé un exemple basique montrant que les SVM à marges douces (CSVM) ne sont pas tolérantes au bruit de classification uniforme, nous proposons une version modifiée de CSVM basée sur une fonction objectif utilisant un estimateur des slack variables du problème non bruité. Les bonnes propriétés de cet estimateur sont appuyées par une analyse théorique ainsi que par des simulations numériques effectuées sur un jeu de données synthétique.

Mots-clés : Classification Supervisée, Séparateurs linéaires, Machines à Vecteurs de Support, Bruit de Classification Uniforme

1 Introduction

Learning from noisy data is a problem of interest both from the practical and theoretical points of view. In this paper, where we address the problem of supervised binary classification, we focus on a particular noise setting where the noise process uniformly flips the true label of an example to the opposite label. This noise, referred to as *uniform classification noise*, was introduced by (Angluin & Laird, 1988).

Some learning algorithm families, such as statistical queries learning algorithms, can be adapted to deal with classification noise. But it is unclear whether the best known methods such as kernel methods can handle such data. For instance, despite soft-margin Support Vector Machines (CSVM) seem to be a viable strategy to learn from noisy data, we show that there are (simple) distributions for which they may fail to learn a good classifier when provided with such data. Here, we propose a noise-tolerant large margin learning algorithm that generalizes CSVMs such that (a) the objective function of the learning algorithm takes into account an unbiased estimation of the non noisy slack errors and, (b) this objective function reduces to the usual CSVM objective when the data are clean. In addition, we show that minimizing this objective function allows to minimize noise-free CSVM objective function if a sufficient number of data is available.

The paper is organized as follows. Section 2 introduces the notations that will be used in the paper. Section 3 shows that CSVM may fail to learn from noisy data. Section 4 presents our noise tolerant version of CSVMs together with its theoretical justifications. Section 5 discusses approaches to learn from noisy data related to the work presented

here. Numerical simulations are presented in Section 6: they show the behavior of our algorithm on a linearly separable distribution in the noise-free and noisy contexts.

2 Notation

$\mathcal{X}$ denotes the input space, assumed to be an *Hilbert space*, equipped with an inner product denoted by $\cdot$. We restrict our study to the binary classification problem and the target space $\mathcal{Y}$ is $\{-1, +1\}$ and the class of functions we focus on is that of hyperplanes from $\mathcal{X}$. These assumptions make our analysis seemlessly applicable with kernels.

Given a fixed but unknown distribution D on the product space $\mathcal{X} \times \mathcal{Y}$, a noise-free sample is a sample $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of data independently and identically distributed according to D. A uniform classification noise process with rate $\eta \in [0, 0.5)$ corrupts a sample $\mathcal{S}$ by independently flipping each y_i to $-y_i$ with probability η. This can be modeled by introducing a *biased*[1] Rademacher vector $\boldsymbol{\sigma}$ of size n such that $\mathbb{P}(\sigma_i = 1) = 1 - \eta$ and $\mathbb{P}(\sigma_i = -1) = \eta$ and saying that $\mathcal{S}^{\boldsymbol{\sigma}} = (\mathbf{x}_i, \sigma_i y_i)_{i=1}^n$ is the noisy version of $\mathcal{S}$. This modeling will be useful in the analysis of our algorithm.

For a linear classifier $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$, the class predicted for $\mathbf{x}$ by f is given by sign $(\mathbf{x})$; depending on the context, f will denote either the linear function or its associated classifier. The *functional* margin $\gamma : \mathbb{R}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is such that $\gamma(f, \mathbf{x}, y) = yf(\mathbf{x}) = y(\mathbf{w} \cdot \mathbf{x} + b)$. For a (noisy) sample $\mathcal{S}$ ($\mathcal{S}^{\boldsymbol{\sigma}}$) and parameters $\mathbf{w}$ and b, we introduce the notation $\gamma_i = y_i(\mathbf{w} \cdot \mathbf{x}_i + b)$ ($\gamma_i^{\boldsymbol{\sigma}} = \sigma_i \gamma_i = \sigma_i y_i(\mathbf{w} \cdot \mathbf{x}_i + b)$) – the dependence of γ_i ($\gamma_i^{\boldsymbol{\sigma}}$) on $\mathbf{w}$ and b is not shown explicitly in the notation since it will always be clear what parameters $\mathbf{w}$ and b are referred to.

The problem that we address in this paper is that of learning a large margin separating hyperplane from $\mathcal{S}^{\boldsymbol{\sigma}} = (\mathbf{x}_i, \sigma_i y_i)_{i=1}^n$, where large margin must be understood as large margin with respect to the true (noise-free) sample.

3 Failure of CSVM on Noisy Data

A very natural thought about the problem that we tackle is that soft margins SVM, in particular L_1-soft margins SVM, on which we will focus, should be tolerant to classification noise. Indeed, recalling that the soft-margin problem applied on a sample $\mathcal{S}$ writes (see, e.g., Schölkopf & Smola (2002))

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + \frac{C}{n} \sum_{i=1}^{n} \ell(\gamma_i), \tag{1}$$

where ℓ is the hinge loss function such that $\ell(\gamma) = 1 - \gamma$ if $1 - \gamma \geq 0$ and 0 otherwise, it might be hoped that the second term of this objective could help dealing with uniform classification noise provided $C > 0$ is well chosen. In other terms, one may think that

[1] $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 0.5$ for a Rademacher variable.

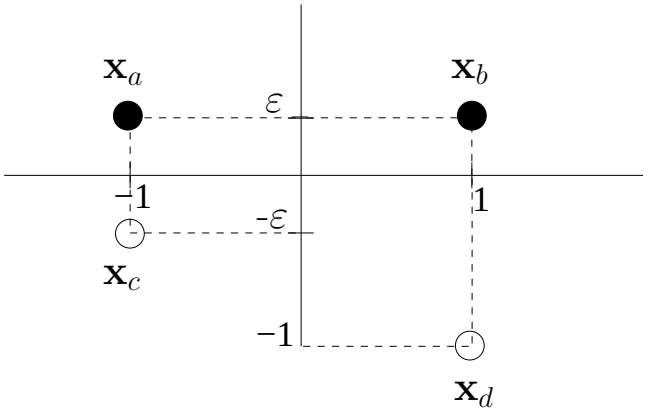

Figure 1: The two-margin distribution $D_{\text{2-margin}}$. Black circles denote positive data and white ones negative data (see Definition 1 for details).

given a noisy sample $\mathcal{S}^{\boldsymbol{\sigma}}$, there exists a value C^* of C such that the solution of

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + \frac{C^*}{n}\sum_{i=1}^{n}\ell(\gamma_i^{\boldsymbol{\sigma}})$$

allows for a generalization error that decreases towards 0 when the number of data grows, if D is linearly separable.

It turns out that it is not the case, i.e. there exist distributions that will make CSVM fail to produce a reliable classifier when applied to a noisy sample. To show that, let us introduce a distribution of data that we call the *two-margin distribution*.

Definition 1 (Two-margin distribution)
We define this distribution $D_{\text{2-margin}}$ on $\mathcal{X}\times\mathcal{Y}$, where $\mathcal{X}=\mathbb{R}^2$. Let $\varepsilon>0$. Let the four points $\mathbf{x}_a=[-1\ \varepsilon]$, $\mathbf{x}_b=[1\ \varepsilon]$, $\mathbf{x}_c=[-1\ -\varepsilon]$ and $\mathbf{x}_d=[1\ -1]$ and the associated classes $y_a=y_b=+1$ and $y_c=y_d=-1$.

Given $p_a,p_b,p_c,p_d>0$ verifying $p_a+p_b+p_c+p_d=1$, an instance drawn according to $D_{\text{2-margin}}$ is such that $\mathbb{P}\left[(\mathbf{x},\mathbf{y})=(\mathbf{x}_a,y_a)\right]=p_a$, $\mathbb{P}\left[(\mathbf{x},\mathbf{y})=(\mathbf{x}_b,y_b)\right]=p_b$, $\mathbb{P}\left[(\mathbf{x},\mathbf{y})=(\mathbf{x}_c,y_c)\right]=p_c$ and $\mathbb{P}\left[(\mathbf{x},\mathbf{y})=(\mathbf{x}_d,y_d)\right]=p_d$. $\mathbf{x}_a$ and $\mathbf{x}_b$ are therefore the locations of the positive instances according to $D_{\text{2-margin}}$ while $\mathbf{x}_c$ and $\mathbf{x}_d$ are the locations of the negative ones.

Figure 1 depicts $D_{\text{2-margin}}$.

We can note that for $\varepsilon>0$, $D_{\text{2-margin}}$ is a linearly separable distribution. In addition, we can sense that if a uniform classification noise corrupts the distribution, a classical CSVM may favor a vertical classification over a horizontal one for some values of $\eta,\varepsilon,p_a,p_b,p_c$ and p_d. This is what is going to be used to proof the following proposition. We first introduce a notion of noise tolerance. Note that this definition is different from that provided by Angluin in the sense that we are interested by consistency on a finite sample.

Definition 2 (Uniform classification noise tolerance)
Let $\mathcal{A}$ be a learning algorithm. $\mathcal{A}$ is said to be tolerant to uniform classification noise, if for all linearly separable distributions D on $\mathcal{X} \times \mathcal{Y}$, for all noise rates $\eta < \frac{1}{2}$, for all $\delta < 1$, there exists $N \in \mathbb{N}$ such that for all S drawn from D of size $n > N$, if $\mathcal{A}$ is given access to $S^{\boldsymbol{\sigma}}$ then, with probability at least $1 - \delta$, $\mathcal{A}$ outputs a classifier consistent with S.

Proposition 1 ()
CSVM *is not tolerant to uniform classification noise: There exists a linearly separable distribution D on $\mathcal{X} \times \mathcal{Y}$ and a noise level $\eta < \frac{1}{2}$ such that, for all $\delta < 1$, there exists $N \in \mathbb{N}$ such that for all S drawn from D of size $n > N$, if* CSVM *is given access to $S^{\boldsymbol{\sigma}}$, then with probability at least $1 - \delta$,* CSVM *outputs a classifier non-consistent with S.*

Proof. In order to prove this proposition, it suffices to show that there exists at least one distribution that makes CSVM not consistent with the noise free version S of $S^{\boldsymbol{\sigma}}$ if the number of data is sufficiently high. ...

Since the proof of this proposition is rather tedious, much part of it is deferred to the appendix (section 8.1). Here, we only give an informal proof, which works in two steps:

1. we assume that CSVM has direct access to a noisy version of distribution $D_{\text{2-margin}}$; optimization problem (1) can therefore be written in terms of the noise rate η and the parameters of $D_{\text{2-margin}}$;

2. we propose a set of values for $\varepsilon, p_a, p_b, p_c, p_d$ and η such that, whatever the value of C in (1), the objective function of (1) for any zero-error hypothesis $\mathbf{w}$ is always larger than the value of the objective function for another hypothesis $\mathbf{w}^\eta$ known to have error at least $\min\{p_a, p_b, p_c, p_d\}$; this simply says that it is not possible to get a zero error hypothesis on this simple linearly separable distribution when corrupted by noise and ends the proof.

3. using Chernoff bounds arguments, it is obvious that a finite sample version can be derived; this ends the proof.

Said otherwise, CSVM may favor the wrong margin if the amount of noise and the value of ε are such that this is the better way to minimize (1). □

4 Proposed Approach

From the analysis of the previous section it turns out that the problem of running the classical CSVM on noisy data comes from the evaluation of the slack errors, accounted for by the second term of (1). If it was possible to estimate the value of the noise free slack errors, it would be possible to accurately learn a large margin classifier from the noisy data. In this section, we show that such an estimation is possible.

4.1 NSVM: a Noise Tolerant Version of CSVM

For a given noise rate η, introduce the following function:

$$\hat{\ell}(\gamma) = \frac{1}{1-2\eta}\left[(1-\eta)\ell(\gamma) - \eta\ell(-\gamma)\right]. \tag{2}$$

We have the following lemma:

Lemma 1
$\forall i \in \{1,\ldots,n\},\ \hat{\ell}(\gamma_i^{\boldsymbol{\sigma}})$ *is an estimator of* $\ell(\gamma_i)$*, i.e.:*

$$\mathbb{E}_{\boldsymbol{\sigma}}(\hat{\ell}(\gamma_i^{\boldsymbol{\sigma}})) = \ell(\gamma_i).$$

Consequently,

$$\mathbb{E}_{\boldsymbol{\sigma}}(\frac{1}{n}\sum_{i=1}^{n}\hat{\ell}(\gamma_i^{\boldsymbol{\sigma}})) = \frac{1}{n}\sum_{i=1}^{n}\ell(\gamma_i)$$

Proof. The proof is straightforward.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}}(\hat{\ell}(\gamma_i^{\boldsymbol{\sigma}})) &= \mathbb{E}_{\sigma_i}(\hat{\ell}(\gamma_i^{\boldsymbol{\sigma}})) \\
&= \frac{1}{1-2\eta}\left[(1-\eta)\mathbb{E}_{\sigma_i}\ell(\gamma_i^{\boldsymbol{\sigma}}) - \eta\mathbb{E}_{\sigma_i}\ell(\gamma_i^{\boldsymbol{\sigma}})\right] \\
&= \frac{1}{1-2\eta}\left[(1-\eta)\left((1-\eta)\ell(\gamma_i) + \eta\ell(-\gamma_i)\right)\right. \\
&\quad \left.-\eta\left((1-\eta)\ell(-\gamma_i) + \eta\ell(\gamma_i)\right)\right] \\
&= \frac{1}{1-2\eta}\left[(1-\eta)^2\ell(\gamma_i) + 0 - \eta^2\ell(\gamma_i)\right] = \ell(\gamma_i).
\end{aligned}$$

□ This lemma simply says that, for given $\mathbf{w}$ and b, we can estimate the actual slack errors from the noisy data. Using $\hat{\ell}$ we may propose a new version of CSVM based on noisy data:

$$\min_{\mathbf{w},b}\frac{1}{2}\mathbf{w}\cdot\mathbf{w} + \frac{C}{n}\sum_{i=1}^{n}\hat{\ell}(\gamma_i^{\boldsymbol{\sigma}}) \tag{3}$$

that we call NSVM (for noisy-SVM). We can note that:

- if the noise rate is $\eta = 0$ then problem (3) comes down to (1);
- the expectation of the objective function of (3) with respect to the noise process is the objective function of (1);
- despite the objective function of (1) is convex, which is an interesting property for optimization purposes, that of (3) is not necessarily convex, because of the non convexity of $\hat{\ell}$ (see (2) and Figure 2).

The next section is devoted to the theoretical analysis of the properties of the objective function of (3), and more precisely to its solution.

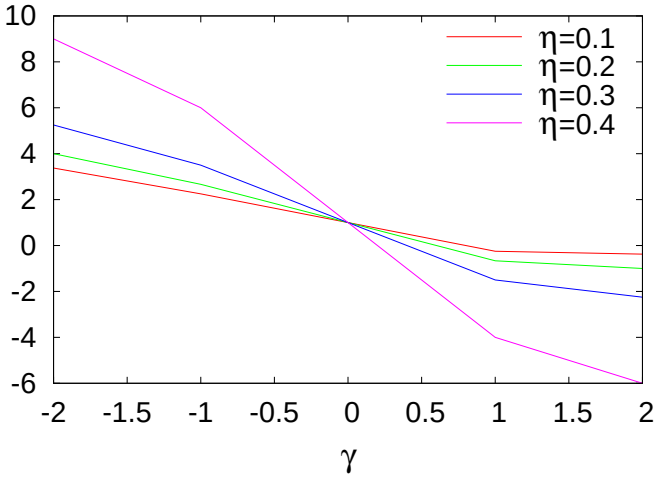

Figure 2: $\hat{\ell}$ as a function of γ for different values of η (see (2)); this function is not convex and is $1/1-2\eta$ -Lipschitz as its largest slope has an absolute value of $1/1-2\eta$.

4.2 Analysis of the Solution of NSVM

Here, the question we address is whether the solution of (3) is close to that of the solution of the corresponding noise free CSVM problem (for the same value of C and the same instances $\mathbf{x}_i$). We show using concentration inequalities and the convex nature of the objective function of (1) that, provided a sufficient number of data is available it is possible to make the solution of (3) arbitrarily close to that of (1).

In order to simplify the proofs, we assume that we are looking for a zero-bias hyperplane, i.e. a separating hyperplane of the form $\mathbf{w} \cdot \mathbf{x} = 0$, and that there exists $R > 0$ such that all $\mathbf{x}$'s drawn from D verify $\mathbf{x} \cdot \mathbf{x} \leq R^2$. Finally, we consider the equivalent Ivanov regularization (Ivanov, 1976; Tikhonov & Arsenin, 1977) forms of (1) and (3), that is, we investigate the closeness of the solutions of:

$$\begin{cases} \min_{\mathbf{w}} G(\mathbf{w}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \ell(\gamma_i) & \text{subject to } \|\mathbf{w}\| \leq W \\ \min_{\mathbf{w}} \hat{G}(\mathbf{w}) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \hat{\ell}(\gamma_i^{\boldsymbol{\sigma}}) & \text{subject to } \|\mathbf{w}\| \leq W \end{cases} \tag{4}$$

for some $W > 0$.

We introduce the notation

$$\mu(\mathcal{S}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\gamma}_i),\ \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\ell}(\gamma_i^{\boldsymbol{\sigma}}), \tag{5}$$

and make use of the following concentration inequality by (McDiarmid, 1989) to establish Lemma 2.

Theorem 1 ((McDiarmid, 1989))
Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $\mathcal{X}$, and assume that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies

$$\sup_{\substack{\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathcal{X} \\ \mathbf{x}_i' \in \mathcal{X}}} \left| f(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n) - f(\mathbf{x}_1, \ldots, \mathbf{x}_i', \ldots, \mathbf{x}_n) \right| \leq c_i$$

for every $1 \leq i \leq n$. Then, for every $t > 0$,

$$\mathbb{P}\left\{ |f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)| \geq t \right\} \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2} \right).$$

Lemma 2
For all distributions D on $\mathcal{X} \times \mathcal{Y}$, $\forall \eta \in [0, 0.5)$, $\forall \delta \in (0, 1]$, $\forall \varepsilon \in \mathbb{R}^+$, for all random sample $\mathcal{S}$ of n examples drawn from D and for all noise vector (of rate η) $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_n\}$, if $n > \frac{8(1+RW)^2}{(1-2\eta)^2\varepsilon^2} \ln \frac{2}{\delta}$ then, with probability at least $1 - \delta$,

$$|\mu(\mathcal{S}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma})| < \varepsilon, \quad \forall \mathbf{w} \in \mathcal{X}, \|\mathbf{w}\| \leq W.$$

Proof. In order to prove the result, it suffices to establish a uniform (wrt $\mathbf{w}$ with $\|\mathbf{w}\| \leq W$) bound on $|\mathbb{E}_{\mathcal{S}}\mu(\mathcal{S}) - \mu(\mathcal{S})|$ and on $|\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma})|$, which is equal to $|\mathbb{E}_{\mathcal{S}}\mu(\mathcal{S}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma})|$. It turns out that an adequate sample size for the latter expression to be lower than $\varepsilon > 0$ is sufficient for the former expression to be lower than ε as well (we let the reader check that). We therefore focus on bounding the function Δ defined as

$$\Delta(\mathcal{S}, \boldsymbol{\sigma}) = \sup_{\mathbf{w} \in \mathcal{X}, \|\mathbf{w}\| \leq W} \left| \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}) \right|.$$

Since $\|\mathbf{w}\| \leq W$ and $\|\mathbf{x}\| \leq R$, the minimum and maximum achievable margin by classifier $\mathbf{w}$ on any instance $(\mathbf{x}, y)$ are $\gamma_{\min} = -RW$ and $\gamma_{\max} = RW$, respectively. Hence, according to the way $\hat{\ell}$ is defined (cf. (2) and Figure 2), it takes values in the range $[\hat{\ell}(\gamma_{\max}); \hat{\ell}(\gamma_{\min})] \subseteq \left[-\frac{\eta(1+RW)}{1-2\eta}; \frac{(1-\eta)(1+RW)}{1-2\eta} \right]$; therefore, the maximum variation of $\hat{\ell}(\gamma_i^{\boldsymbol{\sigma}})$ when changing any $\gamma_i^{\boldsymbol{\sigma}} = \sigma_i\gamma_i$ is at most $\frac{1+RW}{1-2\eta}$. Consequently, the maximum variation of $\Delta(\mathcal{S}, \boldsymbol{\sigma})$ when changing any $\gamma_i^{\boldsymbol{\sigma}}$ is at most $\frac{1+RW}{n(1-2\eta)}$ and, using Theorem 1, we have

$$\mathbb{P}\left\{ |\Delta(\mathcal{S}, \boldsymbol{\sigma}) - \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\Delta(\mathcal{S}, \boldsymbol{\sigma})| \geq \frac{\varepsilon}{4} \right\} \leq 2\exp\left(-\frac{(1-2\eta)^2 n\varepsilon^2}{8(1+RW)^2} \right),$$

which is upper bounded by $\delta/2$ for the choice of n stated in the lemma.

Additionally, we have the following upper bounding on $\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\Delta(\mathcal{S}, \boldsymbol{\sigma})$ (we omit the constraint $\mathbf{w} \in \mathcal{X}, \|\mathbf{w}\| \leq W$ for sake of clarity), where $\boldsymbol{\sigma}'$ is a noise vector with

parameter η, $\boldsymbol{\kappa}$ a vector of n Rademacher variables and $\mathcal{S}'$ a random set of size n drawn from D:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\Delta(\mathcal{S},\boldsymbol{\sigma}) = \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \sup \left|\mathbb{E}_{\mathcal{S}'\boldsymbol{\sigma}'}\hat{\mu}(\mathcal{S}',\boldsymbol{\sigma}') - \hat{\mu}(\mathcal{S},\boldsymbol{\sigma})\right| \\
&\leq \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \sup \mathbb{E}_{\mathcal{S}'\boldsymbol{\sigma}'} \left|\hat{\mu}(\mathcal{S}',\boldsymbol{\sigma}') - \hat{\mu}(\mathcal{S},\boldsymbol{\sigma})\right| && \text{(triangle ineq.)} \\
&\leq \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}\mathcal{S}'\boldsymbol{\sigma}'} \sup \left|\hat{\mu}(\mathcal{S}',\boldsymbol{\sigma}') - \hat{\mu}(\mathcal{S},\boldsymbol{\sigma})\right| && \text{(Jensen ineq.)} \\
&= \frac{1}{n}\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}\mathcal{S}'\boldsymbol{\sigma}'} \sup \left|\sum_{i=1}^{n} \hat{\ell}(\sigma_i' y_i' \mathbf{w}\cdot\mathbf{x}_i') - \sum_{i=1}^{n} \hat{\ell}(\sigma_i y_i \mathbf{w}\cdot\mathbf{x}_i)\right| && \text{(cf. (5))} \\
&= \frac{1}{n}\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}\mathcal{S}'\boldsymbol{\sigma}'\boldsymbol{\kappa}} \sup \left|\sum_{i=1}^{n} \kappa_i \left(\hat{\ell}(\sigma_i' y_i' \mathbf{w}\cdot\mathbf{x}_i') - \hat{\ell}(\sigma_i y_i \mathbf{w}\cdot\mathbf{x}_i)\right)\right| && (\sigma_i y_i \mathbf{x}_i \text{ and } \sigma_i' y_i' \mathbf{x}_i' \text{ are i.i.d}) \\
&\leq \frac{2}{n}\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}\boldsymbol{\kappa}} \sup \left|\sum_{i=1}^{n} \kappa_i \hat{\ell}(\sigma_i y_i \mathbf{w}\cdot\mathbf{x}_i)\right| && \text{(triangle ineq.)} \\
&\leq \frac{2}{n}\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}\boldsymbol{\kappa}} \sup \left|\sum_{i=1}^{n} \kappa_i (\hat{\ell}(\sigma_i y_i \mathbf{w}\cdot\mathbf{x}_i) - 1)\right| + \frac{2}{n}\mathbb{E}_{\boldsymbol{\kappa}}\left|\sum_{i=1}^{n}\kappa_i\right| && \text{(triangle ineq.)} \\
&\leq \frac{2}{n}\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}\boldsymbol{\kappa}} \sup \left|\sum_{i=1}^{n} \kappa_i (\hat{\ell}(\sigma_i y_i \mathbf{w}\cdot\mathbf{x}_i) - 1)\right| + \frac{2}{\sqrt{n}} && \text{(see Appendix 8.2)}
\end{aligned}
$$

We note that $\hat{\ell}() - 1$ is $1/1-2\eta$-Lipschitz and is equal to 0 when its argument is 0. The first term of the last inequality is the Rademacher complexity of the class of functions defined by the composition of $\hat{\ell}() - 1$ and the set of functions defined by zero-bias separating hyperplanes $\mathbf{w}$ such that $\|\mathbf{w}\| \leq W$. The Rademacher complexity of this latter class is bounded from above by $\frac{2WR}{\sqrt{n}}$ (see (Bartlett & Mendelson, 2002)).Using structural results on the Rademacher complexity of composition of functions (see Theorem 12 in (Bartlett & Mendelson, 2002)), we get:

$$\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\Delta(\mathcal{S},\boldsymbol{\sigma}) \leq \frac{4WR}{(1-2\eta)\sqrt{n}} + \frac{2}{\sqrt{n}} \leq \frac{4}{(1-2\eta)\sqrt{n}}(1+WR),$$

which, for the value of n stated in the lemma is upper bounded by $\varepsilon/4$. Therefore, with probability at least $1-\delta/2$ the following holds uniformly over $\mathbf{w}$ for $\|\mathbf{w}\| \leq W$:

$$|\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\hat{\mu}(\mathcal{S},\boldsymbol{\sigma}) - \hat{\mu}(\mathcal{S},\boldsymbol{\sigma})| \leq \frac{\varepsilon}{4} + \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\Delta(\mathcal{S},\boldsymbol{\sigma}) \leq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}.$$

and $|\mathbb{E}_{\mathcal{S}}\mu(\mathcal{S}) - \mu(\mathcal{S})| \leq \varepsilon/2$ with probability $1-\delta/2$ as well. This concludes the proof. □

This lemma says that minimizing $\hat{G}(\mathbf{w})$ as in (4) yields a solution $\mathbf{w}^{\boldsymbol{\sigma}}$ such that the noise free objective $G(\mathbf{w}^{\boldsymbol{\sigma}})$ is close to its minimal value.

Lemma 3
With the same assumptions as in Lemma 2, if $\mathbf{w}^*$ *is the solution of* CSVM *and* $\mathbf{w}^{\sigma}$ *the solution of* NSVM *(run respectively on* $\mathcal{S}$ *and its noisy version* $\mathcal{S}^{\boldsymbol{\sigma}}$*) then, with probability at least* $1-\delta$*,*

$$0 \leq G(\mathbf{w}^{\boldsymbol{\sigma}}) - G(\mathbf{w}^*) \leq \varepsilon.$$

Proof. It suffices to use the fact that $0 \leq G(\mathbf{w}^{\boldsymbol{\sigma}}) - G(\mathbf{w}^*)$ and, with probability $1-\delta$:

$$\begin{aligned} G(\mathbf{w}^{\boldsymbol{\sigma}}) - G(\mathbf{w}^*) &= G(\mathbf{w}^{\boldsymbol{\sigma}}) - \hat{G}(\mathbf{w}^{\boldsymbol{\sigma}}) + \hat{G}(\mathbf{w}^{\boldsymbol{\sigma}}) - \hat{G}(\mathbf{w}^*) \\ &\quad + \hat{G}(\mathbf{w}^*) - G(\mathbf{w}^*) \\ &\leq \frac{\varepsilon}{2} + 0 + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

□

4.3 Implementation

In order to actually solve the NSVM learning we get back to the Tikhonov formulation (3) which have minimized using the BFGS quasi-Newton minimization procedure directly applied to the primal. As proposed by Chapelle (2007), we used a twice-differentiable approximation L_h of the hinge loss function ℓ, with, for $h > 0$

$$L_h(\gamma) := \begin{cases} 0 & \text{if } \gamma > 1+h \\ \frac{(1+h-\gamma)^2}{4h} & \text{if } |1-\gamma| \leq h \\ 1-\gamma & \text{if } \gamma < 1-h \end{cases} \tag{6}$$

Plugging in this loss function in (3) as a surrogate for ℓ, we therefore end up with an unconstrained minimization problem[2] that is easy to solve.

A nice feature of this formulation is that it allows us to establish a result on the closeness of the solution $\mathbf{w}^{\boldsymbol{\sigma}}$ of NSVM and the solution $\mathbf{w}^*$ of CSVM (when they both make use of L_h). In fact, such a result can be also drawn using the usual hinge loss but its non-differentiability makes it a bit more tedious. We have the following proposition.

Lemma 4
For all distributions D on $\mathcal{X} \times \mathcal{Y}$, $\forall \eta \in [0, 0.5)$, $\forall \delta \in (0,1]$, $\forall \varepsilon \in \mathbb{R}^+$, for all random sample $\mathcal{S}$ of n examples drawn from D and for all noise vector (of rate η) $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_n\}$, there exists $N(\eta, \delta, \varepsilon, C) \in \mathbb{N}$ such that if $n > N$ then, with probability at least $1-\delta$,

$$\|\mathbf{w}^{\boldsymbol{\sigma}} - \mathbf{w}^*\|^2 \leq 2\lambda_{\min}^{-1}(K)\varepsilon,$$

where $\lambda_{min}(K)$ is the lowest strictly positive eigenvalue of the Gram matrix $K = (\mathbf{x}_i \cdot \mathbf{x}_j)_{i,j}$

Proof. We only give a sketch of the proof as many of its parts borrow from the proof of previous lemmas (especially the part establishing the precise value of N in order to take C into account).

We therefore assume that for the choice of N given in the lemma: $0 \leq F(\mathbf{w}^{\boldsymbol{\sigma}}) - F(\mathbf{w}^*) \leq \varepsilon$, where $F(\mathbf{w}) = \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + \frac{C}{n}\sum_{i=1}^n L_h(\gamma_i)$ is the objective function of CSVM when making use of L_h.

Let $\mathbf{d} \in \mathcal{X}$. $\forall s \in \mathbb{R}^+, \exists c \in \mathbb{R}^+$ such that:

$$F(\mathbf{w}^* + s\mathbf{d}) = F(\mathbf{w}^*) + s\mathbf{d}^\top \nabla F(\mathbf{w}^*) + \frac{1}{2}s^2 \mathbf{d}^\top H(\mathbf{w}^* + c\mathbf{d})\mathbf{d}$$

[2] The code is available upon request.

where ∇F is the gradient of F and H its Hessian. Since $\mathbf{w}^*$ is the minimum of F, $\nabla F(\mathbf{w}^*) = 0$ and, a few calculations give that $\mathbf{d}^\top H(\mathbf{w}^* + c\mathbf{d})\mathbf{d} \geq \|\mathbf{d}\|^2 \lambda_{\min}(K), \forall \mathbf{d}, c$. Hence, $F(\mathbf{w}^* + s\mathbf{d}) - F(\mathbf{w}^*) \geq \frac{1}{2} s^2 \|\mathbf{d}\|^2 \lambda_{\min}(K)$ and, since $F(\mathbf{w}^* + s\mathbf{d}) - F(\mathbf{w}^*) \leq \varepsilon$, we have $\mathbf{s}^2 \|\mathbf{d}\|^2 \leq 2\varepsilon \lambda_{\min}^{-1}(K)$. □

5 Related Work

Learning from a noisy sample of data implies that the linear problem at hand may not necessarily be consistent, that is, some underlying linear constraints may contradict others. In that case, the problem at hand boils down to that of finding an approximate solution to a linear program such that a minimal number of constraints are violated, which is know as an NP-hard problem (see, e.g., (Amaldi & Kann, 1996)).

In order to cope with this problem, and adapt the classical perceptron learning rule to render it tolerant to noise classification, one line of approaches has mainly been exploited. It relies on exploiting the statistical regularities in the studied distribution by computing various sample averages; this makes it possible to 'erase' the classification noise. As for Bylander's algorithms (Bylander, 1994, 1998), the other notable contributions are those of (Blum *et al.*, 1996) and (Cohen, 1997). However, they tackle a different aspect of the problem of learning noisy distributions and are more focused on showing that, in finite dimensional spaces, the running time of their algorithms can be lowered to something that depends logarithmically on the inverse of the margin instead of linearly. For a kernel version of noise tolerant classifier, one can also looks at (Stempfel & Ralaivola, 2007).

Perceptron-based approaches are not the only ones introduced to handle noisy problems. (Kalai & Servedio, 2005), derived from a boosting algorithm that proposed (Mansour & McAllester, 2000) which can boost with arbitrarily high accuracy in presence of classification noise has been proposed. The algorithm is tolerant to uniform classification noise but also to more complex noise models, such as malicious noise.

Nevertheless, none of this algorithms addresses the problem of learning a maximal margin classifier.

The maximal margin approach we choose to implement is probably the better known of all: Soft-margin SVM problem. (Steinwart, 2002), has proved that, under specific assumptions, CSVM are tolerant to uniform classification noise without modification. Actually, these noise tolerance results are only true with the use of universal kernels (such as Gaussian) and CSVM could perform really poorly on some linear problems (as shown in section 8.1) or with polynomial kernels.

In addition, the proof of the noise tolerance is only valid in the limit, that is to say for an infinite number of examples in the sample. The behavior of the soft-margin learning procedure with C varying as $n^{-1+\beta}$ is still unclear on a given iid finite sample. The modification of SVM we proposed is not kernel dependent, and the accuracy of the estimation of the objective function can be easily computed on a finite sample.

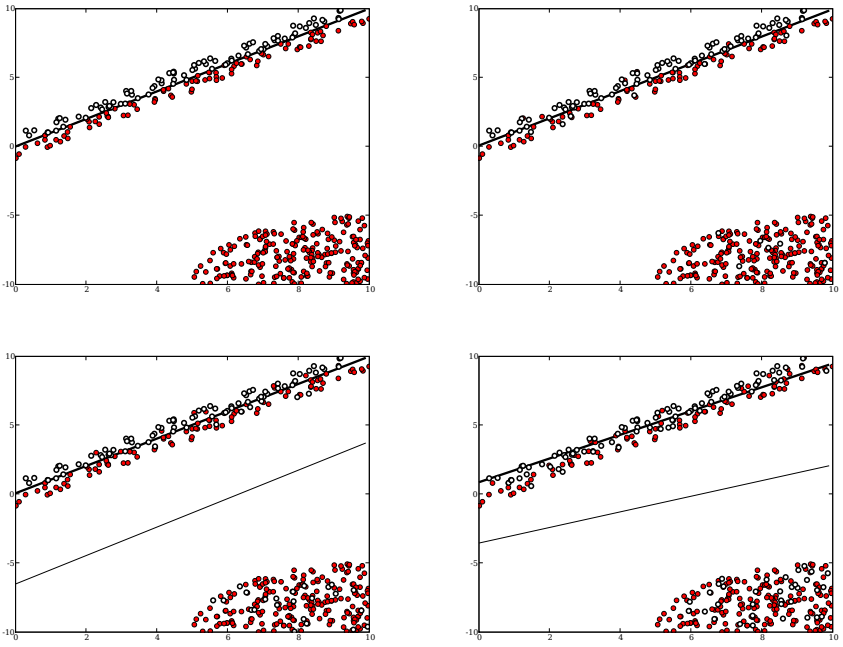

Figure 3: Simulations on a toy problem corrupted by a uniform classification noise rate of 0, 5, 10 and 20 percent. NSVM consistently finds a classifier that makes low error while CSVM is unable to learn good classifier if the noise rate is above 10 percent.

6 Simulations

In order to support our approach, we provide experiments on a linearly separable toy problem. This problem is a derivation of the two-margin distribution (section 3): the sample contains 400 points divided into two clusters separated by a large margin, while the hyperplane used to label the data passes through one of the clusters.

Experiments have been conducted for levels of noise varying from 0 to 20 percent. For each noise rate, we tested 25 values (from 0.0025 to 800000 using a geometric progression) of the parameter C for soft-margin SVM and 5 (from 0.0025 to 25) for our NSVM. The parameter h of the approximate hinge loss was set to $h = 0.1$.

The accuracy of the computed classifiers is estimated on a clean sample of 400 points.

Conclusions of the experiments are quite clear: if the noise rate 5 percent, there exists a parameter C such that CSVM performs well, i.e the classifier output is close to the optimal classifier in the noise-free context. But, as soon as the noise level goes beyond 10 percent, for all the values of C we have tried, the algorithm produces bad classifiers, which make at least 20 percent error. These classifiers define hyperplane that separate the two clusters with the larger margin.

The same experiments made using NSVM exhibit the noise tolerance of our classifier. For noise rates from 0 to 0.20, we obtain, using $C = 2.5$ a classifier with a low

generalization error (even if the sample size is rather small) and which is close from the target classifier. Additional tests have been carried out with NSVM for higher noise rates (from 0.25 to 0.35). With a sample size of 1500, which seems to be a sufficient amount of examples, NSVM still performs well.

7 Conclusion and Outlook

In this paper, we have provided an instance of a linearly separable problem which shows that soft-margin SVM are not naturally tolerant to uniform classification noise. We have proposed a new optimization program, based on CSVM but with a modified objective function. This new objective function makes use of an estimator of slack margin variables in noise-free context. A theoretical analysis proves the noise tolerance of our algorithm, by showing that the solution of NSVM is close to the optimal classifier. Numerical simulations on a noisy problem on which CSVM fails, and on which our algorithm performs well, evidences the good behavior of our learning strategy.

We would like to pursue our research in different directions. First, we will investigate the possibility of finding an other estimator with a lower variance, in order to improve the stability of the algorithm for highly noisy problems or small sample sized problems. In addition, the present estimator is not convex, and this makes more difficult (but still possible and computationally reasonable) the solving of the NSVM problem; It would be of great interest to be able to work out a convex version of the proposed estimator.

8 Appendix

8.1 Proof of Proposition 1

With complete access to distribution D, optimization problem (1) comes down to the minimization of

$$F(\mathbf{w}) = \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + C\mathbb{E}_{(\mathbf{x},y)\sim D}[\ell(y(\mathbf{w}\cdot\mathbf{x}))],$$

which translates, for the noisy version of $D_{\text{2-margin}}$, as

$$F(\mathbf{w}) = \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + C\sum_{i=a,b,c,d} p_i\left[(1-\eta)\ell(\gamma_i) + \eta\ell(-\gamma_i)\right]$$

where $\gamma_i = y_i\mathbf{w}\cdot\mathbf{x}_i$.

Note that when full knowledge of the distribution is available, one should be interested in minimizing F when $C = +\infty$; our analysis holds for this setting when $C \to +\infty$.

Let us assume that $\varepsilon < 0.5$ and that we are looking for a zero-bias separating hyperplane. Note that the maximal margin classifier is of the form $\mathbf{w}^* = [0\ w^*]$ with $w^* \in \mathbb{R}^+$. All zero-error classifiers, i.e classifiers that do not make mistake on the noise-free distribution $D_{\text{2-margin}}$, are of the form $\mathbf{w}_\alpha = [\alpha w\ \ w]$ with $w > 0$ and $\varepsilon > \alpha \geq -\varepsilon$. All other classifiers have error at least $\min\{p_a, p_b, p_c, p_d\}$.

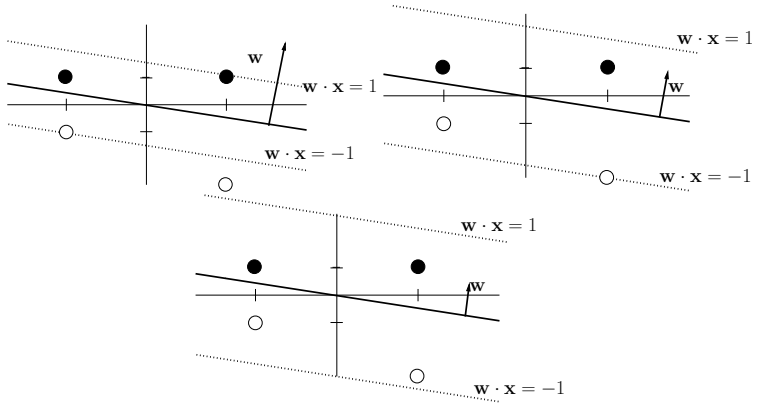

Figure 4: The three cases investigated to proof the non tolerance of CSVM to classification noise. **w**is a vector that classifies correctly all noise-free examples, and its norm varies depending on the considered case.

The remaining of this proof establishes that if $p_a = p_d = 0.4$, $p_c = p_b = 1/2 - p_a = 0.1$, $\varepsilon \in (0, \frac{1}{24}]$ and $\eta = 0.4$ then $\forall C > 0$, $\forall w > 0$, there is a vector $\mathbf{w}' = [-w'\ 0],\ w' > 0$ such that $F(\mathbf{w}') < F(\mathbf{w}_\alpha)$, $\forall \alpha \in [-\varepsilon, \varepsilon)$ and $\mathbf{w}_\alpha = [\alpha w\ w]$; $\mathbf{w}'$ makes error at least $\min\{p_a, p_b, p_c, p_d\}$.

Let $\mathbf{w}' = [-w'\ 0], w' > 0$. Note that, using the definition of the hinge loss function, $\forall w' \in (0, 1]$:

$$\begin{aligned} F(\mathbf{w}') = \frac{w'^2}{2} &+ C(1 + w')\left[(2p_a\eta + (1 - p_a)(1 - \eta)\right] \\ &+ C(1 - w')\left[2p_a(1 - \eta) + (1 - p_a)\eta\right] \end{aligned} \tag{7}$$

and that $\forall w' > 1$,

$$F(\mathbf{w}') = \frac{w'^2}{2} + C(1 + w')\left[2p_a\eta + (1 - p_a)(1 - \eta)\right] \tag{8}$$

For the proposed values of η, p_a, p_b, p_c, p_d, (7) and (8) give $F(\mathbf{w}') = \frac{w'^2}{2} + C(1 - 0.12w')$ and $F(\mathbf{w}') = \frac{w'^2}{2} + C(0.44 + 0.44w')$ respectively.

We investigate three cases on the parameters $\alpha \in [-\varepsilon, \varepsilon)$ and $w > 0$ of a zero-error classifier $\mathbf{w}_\alpha = [\alpha w\ w]$.

First case. Suppose that $w \geq \frac{1}{\varepsilon + |\alpha|}$ (see Figure 4). By definition of $\mathbf{w}_\alpha$, it misclassifies all the noisy points of the distribution, and we have:

$$
\begin{aligned}
F(\mathbf{w}_\alpha) \geq & \frac{w^2(\alpha^2+1)}{2} + C\eta\,[p_a\ell(-\mathbf{w}_\alpha\cdot\mathbf{x}_a) \\
& + p_b\ell(-\mathbf{w}_\alpha\cdot\mathbf{x}_b) + p_c\ell(\mathbf{w}_\alpha\cdot\mathbf{x}_c) + p_d\ell(\mathbf{w}_\alpha\cdot\mathbf{x}_d)] \\
\geq & \frac{w^2}{2} + C\eta\,[p_a(1-(\alpha-\varepsilon)w) + (1/2-p_a)(1+(\varepsilon+\alpha)w) \\
& + (1/2-p_a)(1+(\alpha+\varepsilon)w) + p_a(1-(\alpha-1)w)] \\
= & \frac{w^2}{2} + C\eta\,[1+(1-p_a)\varepsilon w + p_a w + (1-4p_a))\alpha w]
\end{aligned}
$$

And it is straightforward to check that for the values of p_a, η and ε proposed, if $\mathbf{w}' = [-1\ 0]$ then $F(\mathbf{w}') < F(\mathbf{w}_\alpha), \forall w \geq \frac{1}{\varepsilon+|\alpha|}, \alpha \in [-\varepsilon,\varepsilon), C > 0$.

Second case. Suppose that $\frac{1}{1-\alpha} \leq w \leq \frac{1}{\varepsilon+|\alpha|}$. In this case, the classifier errs on the noisy data while the noise-free data located at $\mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c$ are correctly classified, but with a small margin (see Figure 4, middle). Thus

$$
\begin{aligned}
F(\mathbf{w}_\alpha) \geq & \frac{w^2(\alpha^2+1)}{2} + C\eta\,[p_a\ell(-\mathbf{w}_\alpha\cdot\mathbf{x}_a) \\
& + p_b\ell(-\mathbf{w}_\alpha\cdot\mathbf{x}_b) + p_c\ell(\mathbf{w}_\alpha\cdot\mathbf{x}_c) + p_d\ell(\mathbf{w}_\alpha\cdot\mathbf{x}_d)] \\
& + C(1-\eta)\,[p_a\ell(\mathbf{w}_\alpha\cdot\mathbf{x}_a) + p_b\ell(\mathbf{w}_\alpha\cdot\mathbf{x}_b) + p_c\ell(-\mathbf{w}_\alpha\cdot\mathbf{x}_c)] \\
\geq & \frac{w^2}{2} + C\eta\,[1+(1-p_a)\varepsilon w + p_a w + (1-4p_a)\alpha w] \\
& + C(1-\eta)\,[(1-p_a)(1-\varepsilon w) + (3p_a-1)\alpha w] \\
= & \frac{w^2}{2} + C\eta\,[1+(1-p_a)\varepsilon w + p_a w + (1-4p_a)\alpha w] \\
& + C(1-\eta)\,[(1-p_a)(1-\varepsilon w) - (1-p_a)\alpha w]
\end{aligned}
$$

For the proposed values of the parameters, this gives $F(\mathbf{w}_\alpha) \geq \frac{w^2}{2} + C(0.76 + 0.15w)$. If $w \geq 1$, $F(\mathbf{w}_\alpha) > F(\mathbf{w}' = [-1\ 0]) = \frac{1}{2} + C(0.88)$ and $F(\mathbf{w}_\alpha) > F(\mathbf{w}')$. If $w \leq 1$, $F(\mathbf{w}_\alpha) > F(\mathbf{w}' = [-w\ 0]) = \frac{w^2}{2} + C(1-0.12w)$, and, because $\frac{23}{24} \geq \frac{1}{1-\alpha} \geq w$, $F(\mathbf{w}_\alpha) > F(\mathbf{w}')$.

We conclude that $\forall \varepsilon < \frac{1}{24}$, $\forall \alpha \in [-\varepsilon,\varepsilon[$, $\forall w$ such that $\frac{1}{1-\alpha} \leq w \leq \frac{1}{\varepsilon+|\alpha|}$, $\forall C > 0$, there exist $\mathbf{w}'$ such that $F(\mathbf{w}') < F(\mathbf{w}_\alpha)$.

Third case. Suppose that $w \leq \frac{1}{1-\alpha}$. In this third case, all the noisy data are misclassified while the clean data are correctly classified but with a small margin (see Figure 4, right). Thus,

$$
\begin{aligned}
F(\mathbf{w}_\alpha) \geq & \frac{w^2}{2} + C\eta\,[1+(1-p_d)\varepsilon w + p_a w + (1-2(p_a+p_d))\alpha w] \\
& + C(1-\eta)\,[1-(1-p_d)\varepsilon w - p_a w - (1-2(p_a+p_d))\alpha w] \\
= & \frac{w^2}{2} + C \\
& - C(1-2\eta)\,[(1-p_a)\varepsilon w + p_a w + (1-4p_a)\alpha w]
\end{aligned}
$$

For the proposed values of the parameters, this gives $F(\mathbf{w}_\alpha) \geq \frac{w^2}{2} + C(1 - 0.09w)$. If $w \leq 1$, $F(\mathbf{w}_\alpha) > F(\mathbf{w}' = [-w\ 0]) = \frac{w^2}{2} + C(1 - 0.12w)$ and $F(\mathbf{w}_\alpha) > F(\mathbf{w}')$. If $w \geq 1$, $F(\mathbf{w}_\alpha) > F(\mathbf{w}' = [-1\ 0]) = \frac{1}{2} + C(0.88)$, and, because $\frac{25}{24} \geq \frac{1}{1-\alpha} \geq w$, $F(\mathbf{w}_\alpha) > F(\mathbf{w}')$.

We conclude that $\forall \varepsilon < \frac{1}{24}$, $\forall \alpha \in [-\varepsilon, \varepsilon[$, $\forall w$ such that $\frac{1}{1-\alpha} \geq w$,$\forall C > 0$, there exist $\mathbf{w}'$ such that $F(\mathbf{w}') < F(\mathbf{w}_\alpha)$.

It is straightforward, by using Chernoff bounds arguments, to show that, for all $\delta < 1$, there exists $n \in \mathcal{N}$ such that for all samples S drawn from D such that $|S| > n$, the classifier output by CSVM , with access to $S^{\boldsymbol{\sigma}}$,is not consistent with S with probability $1 - \delta$ and, thus, to prove that CSVM is not uniform classification noise tolerant.

Of course, there exist many instances of two-margin problems that would also prevent CSVM classifiers to produce a reliable classifier.

8.2 Bounding $\frac{1}{n}\mathbb{E}_{\boldsymbol{\kappa}} \left|\sum_{i=1}^n \kappa_i\right|$

The proof to bound this quantity is straightforward:

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\kappa}} \left|\sum_{i=1}^n \kappa_i\right| = \frac{1}{n}\mathbb{E}_{\boldsymbol{\kappa}} \sqrt{\sum_{i=1}^n \kappa_i \sum_{j=1}^n \kappa_j} \leq \frac{1}{n}\sqrt{\mathbb{E}_{\boldsymbol{\kappa}} \sum_{i,j=1}^n \kappa_i \kappa_j}$$
$$= \frac{1}{n}\sqrt{\sum_{i=1}^n 1} = \frac{1}{\sqrt{n}}.$$

Where we have used Jensen's inequality (and the concavity of $\sqrt{\cdot}$) in the first line and the fact that $\mathbb{E}_{\kappa_i \kappa_j} \kappa_i \kappa_j = 0$ for $i \neq j$ in the last line.

References

AMALDI E. & KANN V. (1996). On the approximability of some NP-hard minimization problems for linear systems. *Electronic Colloquium on Computational Complexity (ECCC)*, **3**(015).

ANGLUIN D. & LAIRD P. (1988). Learning from Noisy Examples. *Machine Learning*, **2**.

BARTLETT P. L. & MENDELSON S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, **3**, 463–482.

BLUM A., FRIEZE A. M., KANNAN R. & VEMPALA S. (1996). A Polynomial-Time Algorithm for Learning Noisy Linear Threshold Functions. In *Proc. of 37th IEEE Symposium on Foundations of Computer Science*, p. 330–338.

BYLANDER T. (1994). Learning Linear Threshold Functions in the Presence of Classification Noise. In *Proc. of 7th Annual Workshop on Computational Learning Theory*, p. 340–347: ACM Press, New York, NY, 1994.

BYLANDER T. (1998). Learning Noisy Linear Threshold Functions. Submitted for journal publication.

CHAPELLE O. (2007). Training a support vector machine in the primal. *Neural Comput.*, **19**(5), 1155–1178.

COHEN E. (1997). Learning Noisy Perceptrons by a Perceptron in Polynomial Time. In *Proc. of 38th IEEE Symposium on Foundations of Computer Science*, p. 514–523.

IVANOV A. (1976). The theory of approximate methods and their application to the numerical solution of singular integral equations. Nordhoff International.

KALAI A. & SERVEDIO R. (2005). Boosting in the presence of noise. *Journal of Computer and System Sciences*, **71**(3), 266–290.

MANSOUR Y. & MCALLESTER D. (2000). Boosting using branching programs. In *Proc. 13th Annu. Conference on Comput. Learning Theory*, p. 220–224: Morgan Kaufmann, San Francisco.

MCDIARMID C. (1989). On the method of bounded differences. *Survey in Combinatorics*, p. 148–188.

SCHÖLKOPF B. & SMOLA A. J. (2002). *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*. MIT University Press.

STEINWART I. (2002). The influence of the kernel on the consistency of support vector machines.

STEMPFEL G. & RALAIVOLA L. (2007). Learning kernel perceptrons on noisy data using random projections. In *ALT*, p. 328–342.

TIKHONOV A. & ARSENIN V. (1977). Solution of ill-posed problems. Winston, Washington DC.